

A Multivariate Analysis on Non-nucleoside HIV-1 Reverse Transcriptase Inhibitors and Resistance Induced by Mutation

Anna Maria Almerico*, Antonino Lauria, Marco Tutone, Patrizia Diana, Paola Barraja, Alessandra Montalbano, Girolamo Cirrincione and Gaetano Dattolo

Dipartimento Farmacochimico, Tossicologico e Biologico, Università degli Studi, Via Archirafi 32, 90123 Palermo (Italy)

Full Paper

This paper describes the use of multivariate statistical procedure PCA as a tool to explore the inhibitory activity of classes of NNRTIs against HIV-1 viruses (wild type and more frequent mutants, Y181C, V106A, K103N, L100I) and against RT enzyme. The analysis of correlations between biological activity and molecular descriptors or similarity indexes allowed a reliable classification of the fifty five derivatives considered in this study. The best

results were obtained in the case of L100I and K103N mutants for which the higher number of assignments was found when the principal components derived from the descriptors were used. On this basis this statistical approach is proposed as a reliable method for the prediction of the activity of NNRTIs, for which the data against mutant strains have not been reported.

1 Introduction

The treatment regimens for the human immunodeficiency virus (HIV-1) have included both HIV protease and reverse transcriptase (RT) enzyme inhibitors. All antiretroviral drugs currently approved for clinical use are directed against one of these targets. The first RT inhibitors approved in USA and in Europe were nucleoside derivatives (NRTIs) which compete with normal nucleoside substrates for incorporation into the viral genome, thus behaving as chain terminators [1]. Unlike nucleoside analogs, non-nucleoside reverse transcriptase inhibitors (NNRTIs) bind in a non-competitive manner to a specific "pocket" of the HIV-1 RT, which is closely associated with, but distinct from the substrate binding site, altering its ability to function [2].

Nevirapine, Delavirdine, and Efavirenz are the only NNRTIs that have received regulatory approval, whereas several other inhibitors (MKC-442, PNU-142721, and so on) are currently undergoing clinical trials.

Although current NNRTIs have demonstrated potent antiviral activity *in vitro* and in HIV-1 infected patients,

mutations of the reverse transcriptase enzyme associated with NNRTI therapy represent a major problem in developing resistance to current drugs regimens and they limit enormously the effectiveness of the treatment. A single mutation in the NNRTI-binding pocket may result in high-level resistance to one or more NNRTIs. Therefore a high priority for medical research remains the discovery of antiviral agents effective against mutant HIV strains.

In this paper we propose the multivariate statistical procedure, called PCA (principal component analysis), as a tool to exploit the enormous amount of information available on the inhibitory activity against HIV-1 viruses (wild type and the more frequent mutants, Y181C, V106A, K103N, L100I) and against RT enzyme.

Biological problems have an intrinsic multivariate nature, involving many variables at the same time, and in general the relation between these variables and the biological response is hidden and no useful information can easily be extracted. In order to simplify the data set in a multivariate problem and to obtain an informative picture of the data tendencies, a chemometric multivariate analysis can be used.

In the past multivariate data analysis has been applied in many fields of science demonstrating to be most suitable in handling complex data sets and allowing to investigate relationships among all objects and all variables simultaneously. In particular PCA is able to detect similarities among variables and is used to reduce the number of variable thus preparing the data for further analysis. The easier mathematical way to represent a multivariate prob-

* To receive all correspondence. Tel +390916161606, Fax +390916169999, e-mail: almerico@unipa.it

Key words: NNRTIs, PCA, DA, resistance, mutation

Abbreviations: NNRTIs, non nucleoside reverse transcriptase inhibitors; PCA, principal components analysis; DA, discriminant analysis; hit-rate; chi-square.

lem is to build a matrix relating variables and objects. In our case the objects are the NNRTIs, the variables are selected chemicals descriptors, and the biological response are IC_{50} against RT and susceptibility/resistance data.

Each object is now placed in an n -dimensional space (where n is the number of variables). However is more practical for us work in two or three dimensions. The PCA permits the projection of higher order space in two or three dimensions with a minimal loss of statistic information. The coordinate axes of the original n -order space are rotated until the direction of maximum variance is coincident with one of the rotated axes (called the first principal component axis). The second principal component and so on give the orthogonal direction of the maximum residual variance.

Another interesting aspect in a chemometric multivariate analysis is the possibility of classification as done in the discriminant analysis (DA). The derived classification rule describes a surface which separates the classes and it may be used to predict class membership.

It is our aim to develop a simple but efficient method to evaluate, on the basis of chemical-physical descriptors and structural similarity, new NNRTIs that are less likely to trigger resistance or are effective against mutant HIV strains.

2 Materials and Methods

More than 30 different classes of NNRTIs have been described to date. In this study fifty five NNRTI derivatives, the structures of which are depicted in Figure 1, were utilized. The selection includes most of the derivatives currently present into the database of the National Institute of Allergy and Infectious Diseases (NIAID) [3], and the literature derivatives for which the inhibitory concentration against RT enzyme (IC_{50} , μM) was reported [4]. The most active compounds of each class were considered. Table 1 shows these activity values, to be included in the calculations, together with the available data on resistance/susceptibility to the more frequent RT single mutant strains which might confer NNRTI-resistance, reported in literature [5].

The compounds were divided into two classes of activity: high [(H), (IC_{50} up to $0.095 \mu M$)] and low [(L), ($IC_{50} > 0.095 \mu M$)]. The 3D structures of all the derivatives were constructed and optimized by semiempirical methods (MNDO or PM3). For chiral derivatives on which the stereochemistry of the active form is not known both enantiomers were considered (cf. entries 5/6; 27/28; 33/34; 43/44; 48/49).

A selected set of molecular descriptors, which identify the molecular characteristics that can be related to the biological activity, were calculated (ellipsoidal volume, log P, H-bond donors, H-bond acceptors, total lipole, surface area, heat of formation, E_{LUMO} , E_{HOMO} , total dipole moment, total energy in water) [6] and included in the database (Table 2).

Our selection of structural variables was made regarding several reported x-ray crystallographic structures of HIV-1 NNRT inhibitors complexes which supported the already well acknowledged binding mode of the different classes of NNRTIs. In particular ellipsoidal volume, which is defined by the moments of inertia, and accessible surface area give information about steric properties of a molecule, log P is a typical QSAR variable, related to hydrophobic/hydrophilic profile of the inhibitor and in our case can be related to this kind of interaction in the active site, total lipole is a measure of the lipophilic distribution in a 3D space and it is calculated from the summed atomic log P values, the number of H-bond donors and acceptors give other information about the ability of the inhibitor to stabilize its interaction with the "binding pocket", total dipole gives information about the electronic features, E_{LUMO} and E_{HOMO} are energetic variables that classify the set of the inhibitors in terms of their ability to act as electrophiles and nucleophiles, as expected for the inhibition process. Other energy variables are heat of formation, which classifies the set of the inhibitors in terms of relative thermodynamic stability and is widely used in chemometric studies, and total energy in water which classifies the behaviour of the inhibitors in the physiological solvent.

The whole set of 11 descriptors and 55 compounds was used to perform the PCA. Table 3 reports the matrix of eleven PCs with their composition in terms of original variables, together with the fraction of variance explained, the total fraction of variance explained, and the eigenvalue of the covariance matrix corresponding to each component that is equal to the fraction of variance explained by the number of variable used. The data were standardized by mean/sd (standard deviation). The first four PCs which have eigenvalue > 1 and explaining 73.7% of variance, were selected for further calculations.

In the first PC the variables that have major importance are ellipsoidal volume (+0.39), surface area (+0.45), and total energy in water (−0.46). These variables clearly reflect the importance of steric approach to the binding pocket and the behaviour of the molecule within the cytoplasm. In the second PC electronics features have the greater weight: E_{LUMO} (+0.45), E_{HOMO} (+0.45), total dipole moment (−0.47), reflecting the importance of π stacking during the interaction drug-receptor. Log P and heat of formation have great importance in the third and in the fourth PC (+0.59, +0.44 and +0.38, −0.58 respectively), in the same fashion the number of H-bond donors (−0.41) has importance in the third PC, and E_{LUMO} (+0.47) in the fourth. But since the fraction of variance explained by the third and the fourth PC is rather low, these variables are not surely the most relevant.

Since molecular similarity is one of the most useful tool in computer aided approach to discover molecules which bind to the same receptor site, we also calculated the Carbo similarity indexes [7] with flexible optimization in terms of steric feature (shape) and lipophilic potential [6, 8]. In fact

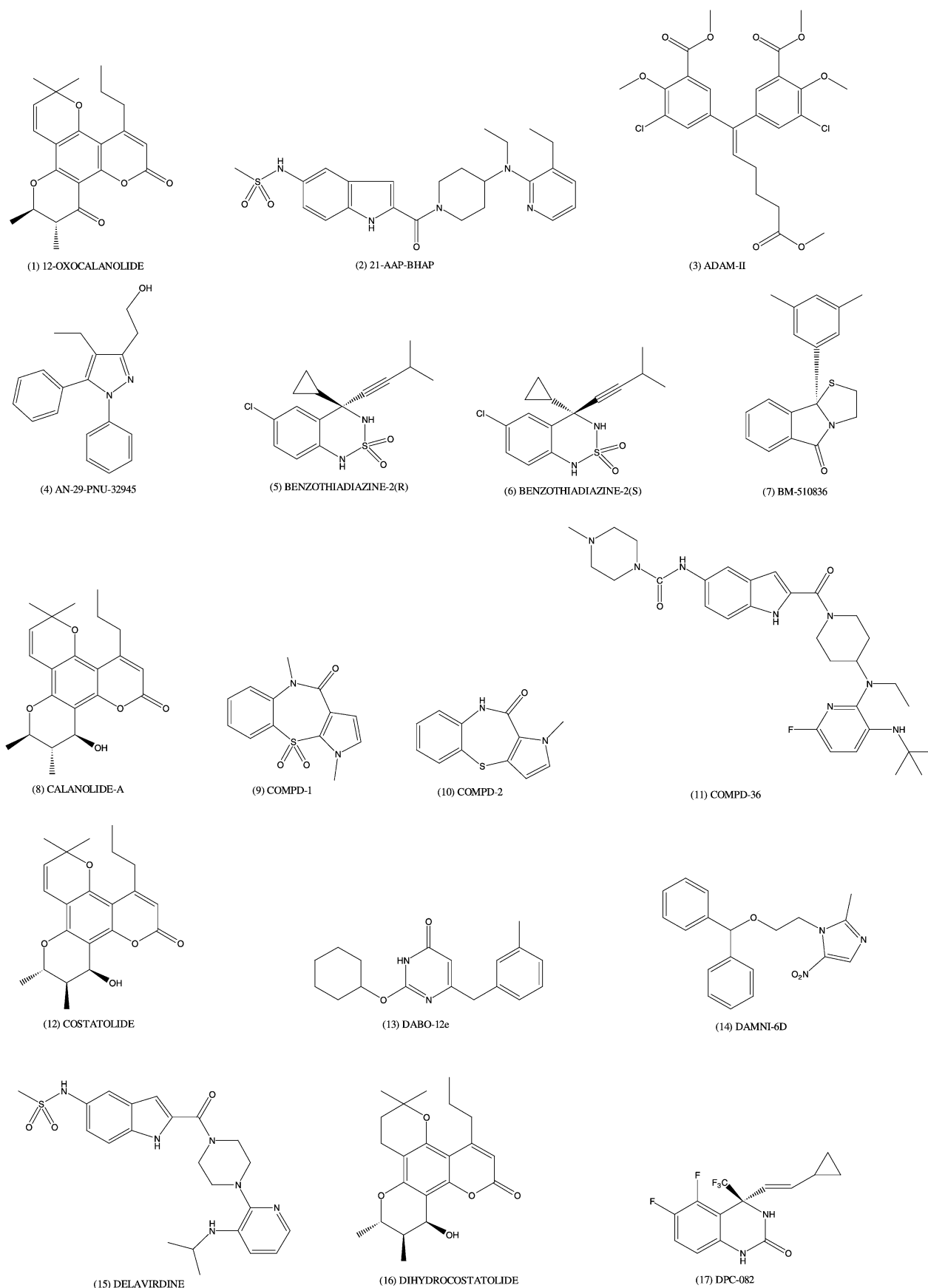


Figure 1.

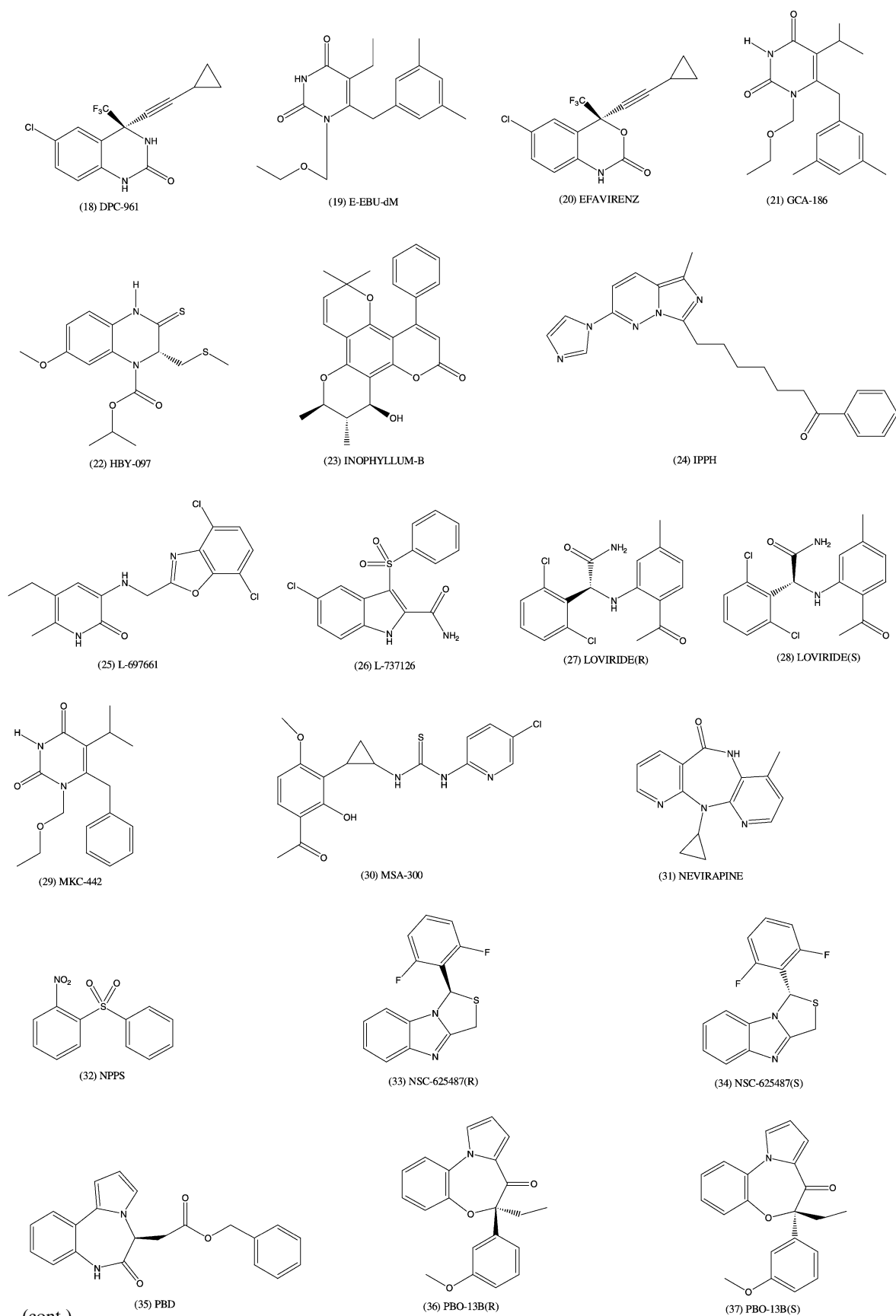


Figure 1. (cont.)

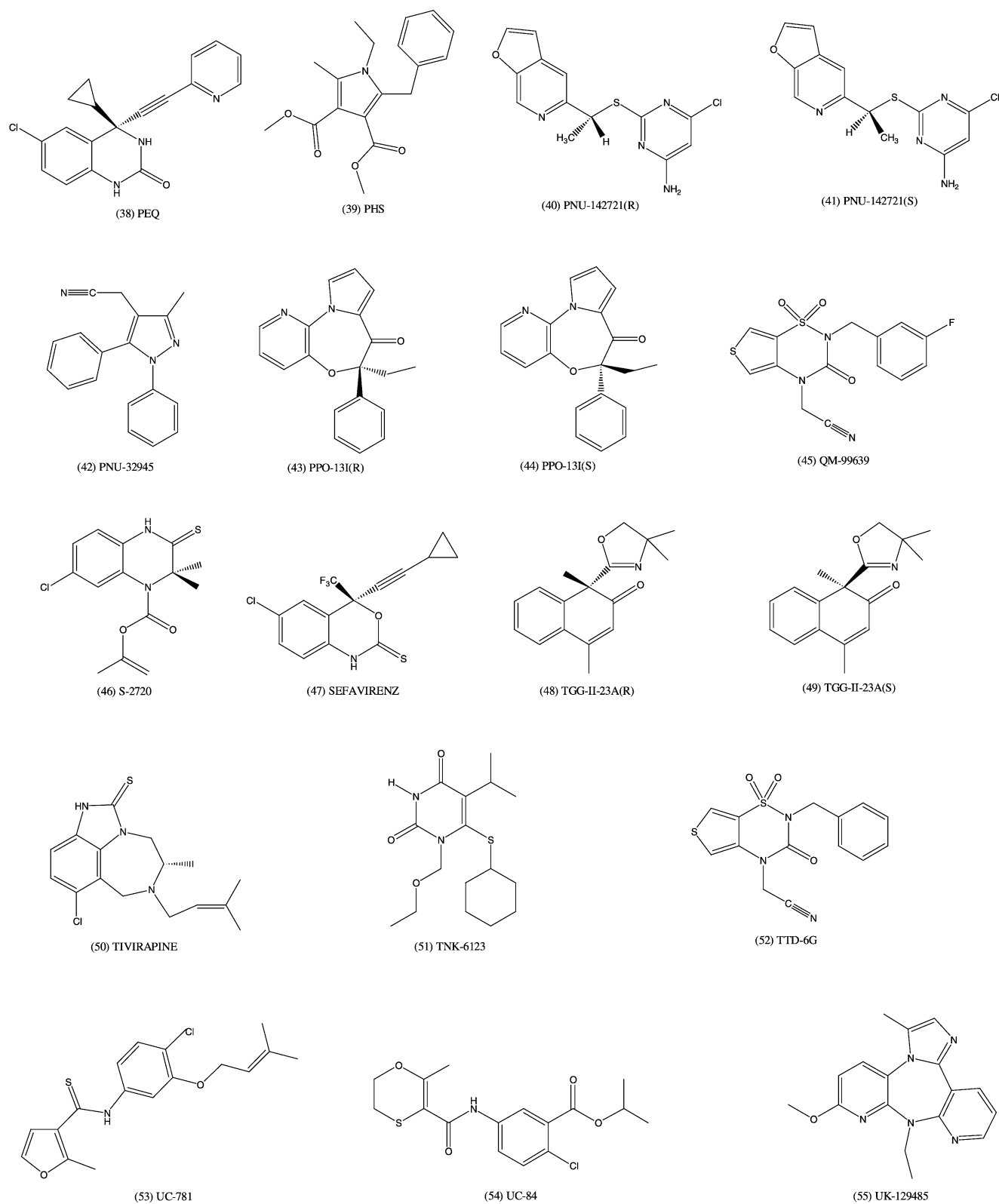


Figure 1. (cont.)

Table 1. RT inhibitory concentrations of NNRTIs and resistance (R)^a/susceptibility (S)^b to selected RT-mutants

Entry	Derivative	IC ₅₀ (μM)	Ref.	L100I	K103N	V106A	Y181C	Activity data	Ref.
1	(+)-12-OXOCALANOLIDE	40	4e	R	R		S	EC ₅₀	3
2	21-AAP-BHAP	0.098	4b				S	IC ₅₀	4b
3	ADAM-II	0.3	4b		R		R	EC ₅₀	4b
4	AN-29-PNU-32945	50	4b						
5	BENZOTHIADIAZINE-2(R)	4.627	4c						
6	BENZOTHIADIAZINE-2(S)	4.627	4c						
7	BM-510836	0.016	4f						
8	CALANOLIDE-A	0.07	4a	R	R		S	EC ₅₀	4a
9	COMP-D-1	1.5	4a						
10	COMP-D-2	28	4a						
11	COMP-D-36	0.21	4b				S	IC ₅₀	4b
12	COSTATOLIDE	0.003	4b	R	R		S	EC ₅₀	3
13	DABO-12e	1.8	4a						
14	DAMNI-6D	0.05	4g						
15	DELAVIRDINE	0.26	4f	S	R	R	R	EC ₅₀	5c,5d
16	DIHYDROCOSTATOLIDE	0.003	4b	R	R		S	EC ₅₀	3
17	DPC-082		4b						
18	DPC-961	0.031	4b		S			EC ₅₀	4b
19	E-EBU-dM	0.036	4f			R		vitro	5a,5b
20	EFAVIRENZ	0.003	4f	R	R	S	S	EC ₅₀	3,5d,5c
21	GCA-186		4b		R		R	EC ₅₀	4f
22	HBV-097	0.08	4f		R			EC ₅₀	4f
23	INOPHYLLUM-B	0.038	4a						
24	IPPH	0.0006	4a						
25	L-697661	0.019	4a	S	S		R	vitro	5a,5b
26	L-737126	0.003	4a						
27	LOVIRIDE(R)	0.2	4a		S		R	vitro	5a,5b
28	LOVIRIDE(S)	0.2	4a		S		R	vitro	5a,5b
29	MKC-442	0.012	4a		R		R	EC ₅₀	4a
30	MSA-300		4a						
31	NEVIRAPINE	0.084	4f	S	R	R	R	EC ₅₀	3
32	NPPS	0.61	4a						
33	NSC-625487(R)	0.5	4a						
34	NSC-625487(S)	0.5	4a						
35	PBD	0.115	4a						
36	PBO-13B(R)		4b	S	S	S		K _i	4b
37	PBO-13B(S)		4b	S	S	S		K _i	4b
38	PEQ	0.012	4a						
39	PHS	0.64	4a						
40	PNU-142721(R)	0.085	4b				R	IC ₅₀	4b
41	PNU-142721(S)	0.02	4b	S			S	IC ₅₀	4b
42	PNU-32945	2.31	4b						
43	PPO-13I(R)		4b	S	S	S		K _i	4b
44	PPO-13I(S)		4b	S	S	S		K _i	4b
45	QM-99639	15.1	4b	S	R		R	EC ₅₀	4b
46	S-2720	0.1	4f			R		vitro	5a,5b
47	SEFAVIRENZ		4b						
48	TGG-II-23A(R)	136	4a						
49	TGG-II-23A(S)	136	4a						
50	TIVIRAPINE	0.05	4a		R			vitro	5a,5b
51	TNK-6123		4b		S		S	EC ₅₀	4b
52	TTD-6G	16.4	4h	S	R		R	EC ₅₀	4 h
53	UC-781	0.059	4a	R	R		S	EC ₅₀	4a
54	UC-84	1.5	4a	R			R	vitro	5a,5b
55	UK-129485	0.156	4a						

^aR > 15 fold resistant ratio (Activity on mutant strain/Activity on wt); ^bS < 15 fold resistant ratio.

Table 2. Calculated molecular descriptors for NNRTI derivatives

Entry	Descriptors										
	1	2	3	4	5	6	7	8	9	10	11
1	628.9	3.2	0	5	5.51	403.8	− 143.82	− 0.859	− 8.922	8.45	− 4748.8
2	3000.0	2.4	2	4	17.72	515.5	− 8.08	− 0.630	− 8.831	5.59	− 5692.5
3	1652.6	4.4	0	8	7.09	547.3	− 282.18	− 0.783	− 9.420	2.42	
4	614.3	4.8	1	2	13.19	351.4	60.09	− 0.160	− 8.878	3.01	− 3440.1
5	763.5	2.8	2	2	9.28	339.5	25.05	− 0.672	− 9.276	3.77	− 3774.8
6	763.5	2.8	2	2	9.28	339.5	25.05	− 0.672	− 9.276	3.77	− 3774.8
7	578.3	4.8	0	1	6.86	322.6	38.24	− 0.599	− 8.728	2.44	− 3259.0
8	654.3	3.3	1	5	5.33	410.9	− 167.61	− 0.763	− 8.826	6.25	− 4776.9
9	351.4	1.0	0	3	4.40	265.5	− 4.28	− 0.706	− 9.479	5.53	− 3396.7
10	300.8	1.5	1	1	6.28	235.0	51.25	− 0.311	− 7.909	2.37	− 2601.4
11	5434.5	2.0	3	4	19.97						
12	654.9	3.3	1	5	5.33	411.3	− 167.61	− 0.764	− 8.827	6.26	− 4776.9
13	691.0	3.5	1	3	8.15	358.2	− 49.34	− 0.230	− 9.245	4.11	− 3665.4
14	743.5	4.3	0	4	13.28	377.3	64.29	− 0.870	− 9.582	4.99	− 4270.9
15	3450.6	0.7	3	4	10.59	492.3	10.13	− 0.636	− 8.559	7.65	− 5603.0
16	623.4	3.4	1	5	4.93	415.5	− 195.20	− 0.705	− 8.932	6.92	− 4805.4
17	638.3	3.0	2	1	6.13	297.7	− 198.04	− 0.803	− 9.318	3.69	− 5102.7
18	735.3	3.0	2	1	7.08	300.7	− 81.49	− 0.529	− 9.196	4.04	− 4491.2
19	839.3	2.8	1	3	11.25	373.6	− 102.35	− 0.190	− 9.374	4.12	− 4011.1
20	674.7	3.6	1	2	5.48	297.6	− 120.15	− 0.749	− 9.454	5.36	− 4590.7
21	875.0	3.1	1	3	11.39	389.5	− 104.31	− 0.185	− 9.370	3.90	− 4166.5
22	1002.9	2.4	1	4	2.10	369.6	− 63.41	− 0.844	− 8.454	5.45	− 3959.0
23	712.4	3.8	1	5	8.93	426.0	− 118.05	− 0.860	− 8.803	6.27	− 5132.1
24	3281.9	4.8	0	4	7.20	455.3	131.06	− 0.888	− 8.641	3.99	− 4650.8
25	1012.3	1.3	2	3	15.68	358.2	− 0.77	− 0.781	− 8.197	3.36	− 4265.1
26	807.7	1.8	2	3	10.77	311.9	− 26.68	− 0.990	− 9.349	3.77	− 4014.4
27	702.8	2.9	2	2	5.29	351.4	− 38.49	− 0.424	− 8.671	2.65	− 4199.5
28	702.8	2.9	2	2	5.29	351.4	− 38.49	− 0.424	− 8.671	2.65	− 4199.5
29	638.1	2.2	1	3	9.49	345.4	− 89.08	− 0.218	− 9.406	3.56	− 3854.8
30	2512.9	3.4	3	5	14.28	396.6	− 3.11	− 0.586	− 8.369	5.42	− 4705.8
31	421.9	1.9	1	3	1.98	287.4	87.90	− 0.564	− 8.902	3.19	− 3258.8
32	601.2	2.8	0	4	7.10	244.5	6.97	− 1.404	− 10.395	9.37	− 3338.6
33	375.9	4.1	0	1	4.69	279.5	29.32	− 0.785	− 8.972	3.21	− 3633.9
34	375.9	4.1	0	1	4.69	279.5	29.32	− 0.785	− 8.972	3.21	− 3633.9
35	1793.4	3.0	1	3	7.48	370.9	− 11.47	− 0.267	− 8.458	1.73	− 4310.4
36	804.1	4.1	0	3	5.20	356.7	10.96	− 0.475	− 8.964	2.31	− 4115.2
37	806.7	4.1	0	3	5.20	356.7	10.96	− 0.476	− 8.964	2.32	− 4115.3
38	876.9	3.3	2	2	9.88	339.8	115.59	− 0.501	− 8.922	2.15	− 3808.9
39	584.6	2.8	0	4	9.14	366.5	− 114.99	0.212	− 8.984	3.80	− 4082.7
40	1636.5	3.2	1	4	3.71	304.8	72.63	− 0.472	− 8.545	5.90	− 3530.7
41	913.8	3.2	1	4	3.65	304.4	72.03	− 0.427	− 8.558	2.15	− 3530.7
42	524.2	4.4	0	2	7.45	318.3	149.51	− 0.279	− 9.099	4.18	− 3128.3
43	612.2	3.5	0	3	4.15	321.7	62.54	− 0.746	− 9.078	3.57	− 3704.0
44	612.2	3.5	0	3	4.15	321.7	62.54	− 0.746	− 9.078	3.57	− 3704.0
45	553.7	1.8	0	4	6.32	309.8	− 32.82	− 1.121	− 9.528	5.65	− 4385.8
46	539.3	2.8	1	3	5.95	322.4	2.74	− 1.015	− 8.747	2.84	− 3620.0
47	737.5	4.7	1	2	4.64	311.9	− 57.06	− 1.070	− 9.230	5.69	− 4463.7
48	494.4	2.9	0	3	6.84	313.8	− 14.20	− 0.746	− 9.285	3.55	− 3284.9
49	494.4	2.9	0	3	6.84	313.8	− 14.20	− 0.746	− 9.285	3.55	− 3284.9
50	682.2	3.8	1	2	3.54	345.9	74.37	− 0.616	− 8.400	3.33	− 3511.2
51	511.3	3.1	1	3	6.92	375.7	− 131.61	− 0.680	− 9.054	3.70	− 3978.7
52	451.3	1.7	0	4	5.96	304.6	12.13	− 1.048	− 9.584	5.35	− 3913.7
53	1420.7	4.6	1	3	4.90	367.4	12.25	− 0.948	− 8.494	5.51	− 3839.4
54	1610.2	1.1	1	4	13.32	370.7	− 128.30	− 0.419	− 8.573	6.54	− 4354.3
55	536.8	3.6	0	4	5.19	334.0	112.75	− 0.599	− 8.495	2.53	− 3762.5

Value were obtained by using numerical algorithms included in TSAR 3.2 package: 1. Ellipsoidal volume(Å³), 2. log P, 3. number of H-bond donors, 4. number of H-bond acceptors, 5. total lipole, 6. surface area (Å²), 7. heat of formation (kcal mol^{−1}), 8. E_{LUMO} (eV), 9. E_{HOMO} (eV), 10. total dipole moment (Debye), 11. total energy (eV).

Table 3. Results of PCA on the reduced set of descriptors and similarity indexes

<i>Descriptor</i>	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
1	0.386	0.238	0.259	−0.319	−0.096	0.052	−0.494	−0.111	0.216	−0.424	0.357
2	−0.140	0.034	0.589	0.438	−0.350	0.338	0.069	0.371	−0.143	−0.195	0.015
3	0.278	0.302	−0.408	−0.150	0.043	0.436	−0.016	0.329	−0.580	0.013	0.052
4	0.331	−0.262	0.272	−0.027	0.257	−0.487	0.232	0.013	−0.517	−0.344	−0.033
5	0.288	0.223	−0.224	−0.083	−0.607	−0.290	0.497	0.172	0.246	−0.128	−0.052
6	0.450	0.124	0.267	0.213	−0.053	−0.067	0.057	−0.191	−0.080	0.707	0.331
7	−0.227	0.281	0.385	−0.581	−0.201	−0.121	−0.084	−0.026	−0.234	0.275	−0.434
8	−0.027	0.454	−0.150	0.473	0.035	−0.485	−0.471	0.171	−0.064	−0.008	−0.228
9	0.094	0.453	0.225	−0.037	0.619	0.095	0.366	0.239	0.365	−0.025	−0.119
10	0.293	−0.467	0.025	−0.153	0.026	−0.053	−0.281	0.655	0.249	0.246	−0.189
11	−0.457	0.075	−0.010	−0.203	0.046	−0.321	0.055	0.396	−0.070	0.092	0.681
Fraction of variance explained	0.327	0.194	0.120	0.095	0.083	0.075	0.031	0.029	0.022	0.010	0.009
Total variance explained	0.327	0.521	0.642	0.737	0.820	0.896	0.927	0.957	0.979	0.991	1
Eigenvalue	3.602	2.137	1.325	1.046	0.917	0.834	0.341	0.326	0.246	0.120	0.100
<i>Shape</i>	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8			
Fraction of variance explained	0.529	0.118	0.091	0.045	0.034	0.029	0.020	0.019			
Total variance explained	0.529	0.647	0.739	0.784	0.819	0.848	0.869	0.889			
Eigenvalue	29.130	6.491	5.050	2.498	1.892	1.608	1.137	1.082			
<i>Lipophilicity</i>	PC1	PC2									
Fraction of variance explained	0.947	0.024									
Total variance explained	0.947	0.971									
Eigenvalue	52.1	1.33									

the shape could provide information on the accessibility and interaction with the active site, whereas the lipophilicity can be important not only for the interaction with the binding site, which is stabilized by hydrophobic features of the inhibitor, but also for the bioavailability.

The measure of similarity between two molecules, based on the physical properties of lipophilicity, shape, electrostatic potential and so on, was obtained by using an analytical approach in which Gaussian functions are substituted into the calculation equations. The properties were compared throughout space, using an infinitesimal grid increment. Molecular similarity indexes are very sensitive to the relative orientation of the two compounds and depend on pair wise molecular superimpositioning. The optimization was achieved by rotation and translation of the molecules for comparison and by altering the conformation of these molecules (flexible optimization).

Again a PCA was carried out on the matrices of similarity indexes and the number of significant PCs was determined as above (eigenvalue > 1). These PCs (8 in the case of similarity index according to shape and 2 in the case of similarity index according to lipophilicity) were considered (cf Table 3).

Therefore the PCs selected for descriptors and similarity indexes were separately included in a discriminant analysis (DA) with the aim of providing assignment to the two classes of activity H or L. By using the Mahalanobis distance discrimination algorithm and the procedure reported by Manly [9], the TSAR automatic procedure was selected as

stopping procedure. This ends when all the points are correctly classified (no variable will give more than 5% increase in the total Mahalanobis distance sum between class centers and the best variable reduces the total number of well-classified points) or all the variables have been added. A data point is considered to be well-classified if the classification rule derived by the DA predicts that point to belong to its true class. A summary of the results achieved with DA is reported in Table 4 that lists the total number of compounds predicted to belong to each class (H or L), broken down by the true class of each individual, the final set of PCs included in each case, together with the partial hit-rate. The mean of the partial hit-rates gives the total hit-rate (percentage of statistical units correctly classified by the discriminant function). This last is an estimate of the true predictive ability of the model which was validated by Huberty's tests [proportional chance (C_{pro}) and maximum chance (C_{max})] [10]. The proportional chance criterion for assessing model fit is calculated by summing the squared proportion that each group represents of the sample (Eq. 1). The maximum chance criterion is the proportion of cases in the largest group (Eq. 2).

$$C_{\text{pro}} = p^2 + (1 - p)^2 \quad (1)$$

p = proportion of subjects in one group

$(1 - p)$ = proportion of cases in the other group

$$C_{\text{max}} = (n_L/N_L) (100) \quad (2)$$

Table 4. Results of DA^a based on PCs for descriptors and similarity indexes for classes **H** and **L**

	Descriptors				Similarity index SHAPE				Similarity index LIPOPHILICITY			
	Predicted		Total	Hit-Rate	Predicted		Total	Hit-Rate	Predicted		Total	Hit-Rate
	L	H			L	H			L	H		
L	18	6	24	75.0%	18	6	24	75.0%	17	9	26	65.3%
H	8	12	20	60.0%	5	14	19	73.6%	7	13	20	65.0%
Total	26	18	44^b	68.1%	23	20	43^c	74.4%	24	22	46^d	65.2%
<i>Included variables</i> PC1, PC2, PC3					PC4, PC6, PC8				PC1, PC2			
Chi2	5.52; df = 1, p < 0.025				10.10; df = 1, p < 0.01				4.18; df = 1, p < 0.05			
Cpro	50.4% min. hit-rate = 63.0%				50.6% min. hit-rate = 63.2%				50.8% min. hit-rate = 63.5%			
Cmax	54.5% min. hit-rate = 64.0%				55.8% min. hit-rate = 69.7%				56.5% min. hit-rate = 70.6%			

^a Cross validation employed: leave out 1 row in turn. ^b Excluded entries: 3, 11, 17, 21, 30, 36, 37, 43, 44, 47, 51. ^c Excluded entries: 11, 14, 17, 21, 30, 36, 37, 39, 43, 44, 47, 51. ^d Excluded entries: 17, 21, 30, 36, 37, 43, 44, 47, 51.

n_L = number of subjects in the larger of the two groups
 N_L = total number of subjects into combined groups

The classification accuracy of the model should result one fourth greater than that achieved by chance (min. hit-rate).

The other validation test we used in this analysis is Pearson's chi-square (Chi2) [11]. The chi-square test of statistical significance is a series of mathematical formulas which compare the actual observed frequencies of some phenomenon (in our sample) with the frequencies we would expect if there were no relationship at all between the two variables in the larger (sampled) population. Therefore, chi-square tests the actual results against the null-hypothesis and assesses whether they are different enough to overcome a certain probability that they are due to sampling error.

3 Results and Discussion

Data reported in Table 4 show the results of the classification obtained when evaluating the inhibitory activity against RT enzyme. After selecting the appropriate number of PCs as reported in the previous section and excluding the entries exhibiting missing values, the discriminant analyses provided a good assignment of the derivatives to the two classes of activity. In particular if shape is considered, the model correctly classifies both high and low active compounds, on the basis of the 8 PCs (explaining 89% of variance). An hit-rate = 74.4% demonstrated the predictivity of the model, whereas in the case of similarity index based on lipophilicity, the hit-rate resulted lower (65.2%). In the case of PCs derived from descriptors, when 4 PCs (eigenvalue > 1, cfr Table 3) are included in the calculations, the less active compounds were well classified (partial hit-rate 75%); but a lower value was obtained in the case of compounds belonging to the H class. However the total hit-rate (68.1%) results statistically significant. In fact based on the requirement for the model accuracy (25% better than Cpro) the standards to use for comparing the model accuracies are 63.0% (descriptors), 63.2% (shape indexes),

63.5% (lipophilicity indexes). Our model accuracy rates exceed these standards. With respect to the Cmax our model accuracy rates exceed the standards in all the cases but for lipophilicity indexes. However all the models present chi-square > 3.84 and p (probability value) < 0.05 [as required when df (degrees of freedom) = 1].

According to the DA, besides entry **31** (Nevirapine), that was always classified as low active probably because the IC_{50} is rather close to the value used for separating the two classes, also for derivatives **50** (Tivirapine) and **53** (UC-781) (in similarity indexes) and **25** (L-697661), in lipophilicity index and descriptors, the misclassification as low active is probably due to their neighbourhood to the cut-off point. The analysis misclassified as high active derivatives **2** (21-AAP-BHAP) for the same reason as above. Only for entries **20** (Efavirenz), **24** (IPPH), **26** (L-737126), the misclassification is not justified by this assumption.

Based on descriptors the misclassification as high active involves entries **15** (Delavirdine), **35** (PBD), **42** (PNU-32945), and **55** (UK-129485), whereas as far as the similarity index based on shape was concerned the same type of misclassification involved derivatives **10** (Compd 2), **15** (Delavirdine), **27/28** (Loviride), and **54** (UC-84). In the analysis based on lipophilicity indexes the entries **2**, **4**, **5/6**, **9**, **10**, **13**, **42** were classified as high active. 12-Oxocalanolide (**1**) was always classified as high active although the IC_{50} was 40 μ M, probably because of the obvious structural similarity with Calanolide (**8**), Costatolide (**12**) and Dihydrocostatolide (**16**) that were all correctly assigned to the expected H class.

Since most NNRTIs bind to the same enzyme's pocket, structural requirements are very strict and it is particularly significant that the best results (higher hit-rate) are obtained when the shape was considered. Therefore the inhibitory activity of this series of NNRTIs can be properly described on the basis of the statistical approach outlined so far.

The next step was to demonstrate that PCA can be used when the activity or inactivity against the RT mutant was considered. To this purpose the available data on the activity

against the more common HIV-1 mutants were considered for calculation. These data include the inhibitory activity measured against the cells infected with mutated virus (EC_{50} values) or against mutated RT enzyme (IC_{50} or K_i values) or other *in vitro* tests (cfr Table 1).

It is well known that a single mutation in the NNRTI-binding pocket may result in high-level resistance to one or more NNRTIs. The RT mutations selected in this paper are K103N and Y181C, associated with resistance to NNRTIs which compromise the antiviral activity of first generation of NNRTIs [4b], including the available drugs [(15) Delavirdine, (31) Nevirapine], and V106A and L100I which cause resistance also to the newer generations of NNRTIs, although from low to intermediate in the case of (15) Delavirdine and (20) Efavirenz [12, 13].

This time the NNRTIs were classified as susceptible (S)/resistant (R) against the single mutant strain [5] and the DA was carried out considering in turn only the derivatives for which the data on resistance/susceptibility were available (cfr Table 1). The results of these discriminant analyses are

summarized in Table 5. The hit-rates obtained in these cases were generally very good (≥ 80.9) especially for the susceptible compounds: correct assignment to the classes was achieved with very few exceptions. These findings were particularly significant in the case of the two more frequent mutants (L100I and K103N), when the PCs derived from descriptors and shape similarity index are considered. When the similarity indexes based on lipophilicity were used the hit-rate had statistic significance only in the case of L100I mutant. Taking in account the PCs derived from descriptors it is relevant to observe that in the case of L100I and K103N mutations, the first two principal components have a major weight evidencing thus the importance of shape and electronics features in the arising resistance. In the case of Y181C mutation again the first PC is mainly involved, but an important role has the third and the fourth PC (log P, heat of formation, but also E_{LUMO} and H-bond donors). In the case of V106A mutation the second, the third, the fourth PCs are relevant so we can assume that electronics features carry out a role of primary importance.

Table 5. Results of DA^a based on PCs for descriptors and similarity indexes for classes **R** and **S**

		Descriptors			Similarity index SHAPE				Similarity index LIPOPHILICITY				
		Predicted		Tot.	Hit Rate	Predicted		Tot.	Hit Rate	Predicted		Tot.	Hit Rate
		R	S			R	S			R	S		
L	R	6	1	7	85.7%	7	0	7	100%	6	1	7	85.7%
	S	0	10	10	100%	0	10	10	100%	1	9	10	90.0%
	Total	6	11	17	94.1%	7	10	17	100%	7	10	17	88.2%
	Included variables PC1, PC2, PC4												
	PC4, PC6, PC7												
0	Chi2	13.24; df = 1, p < 0.001			17; df = 1, p < 0.001				9.74; df = 1, p < 0.01				
I	Cpro	51.5% min. hit-rate = 64.3%			51.5% min. hit-rate = 64.3%				51.5% min. hit-rate = 64.3%				
	Cmax	58.8% min. hit-rate = 73.5%			58.8% min. hit-rate = 73.5%				58.8% min. hit-rate = 73.5%				
K	R	11	4	15	73.3%	14	2	16	87.5%	15	1	16	93.7%
	S	0	8	8	100%	0	8	8	100%	7	1	8	12.5%
	Total	11	12	23 ^b	82.6%	14	10	24	91.6%	22	2	24	66.7%
	Included variables PC1, PC2, PC4												
	PC3, PC4, PC5, PC8												
3	Chi2	11.24; df = 1, p < 0.001			13.24; df = 1, p < 0.001				0.27; df = 1, p < 1				
N	Cpro	52.7% min. hit-rate = 65.8%			55.5% min. hit-rate = 69.7%				55.5% min. hit-rate = 69.7%				
	Cmax	65.2% min. hit-rate = 81.5%			66.7% min. hit-rate = 83.3%				66.7% min. hit-rate = 83.3%				
V	R	4	0	4	100%	4	0	4	100%	2	2	4	50.0%
	S	0	5	5	100%	0	5	5	100%	0	5	5	100%
	Total	4	5	9	100%	4	5	9	100%	2	7	9	77.8%
	Included variables PC2, PC3, PC4												
	PC3												
6	Chi2	9; df = 1, p < 0.01			9; df = 1, p < 0.01				3.21; df = 1, p < 0.1				
A	Cpro	50.5% min. hit-rate = 63.1%			50.5% min. hit-rate = 63.1%				50.5% min. hit-rate = 63.1%				
	Cmax	55.5% min. hit-rate = 69.4%			55.5% min. hit-rate = 69.4%				55.5% min. hit-rate = 69.4%				
Y	R	7	1	8	87.5%	5	3	8	62.5%	7	2	9	77.8%
	S	1	11	12	91.6%	1	12	13	92.3%	7	6	13	46.1%
	Total	8	12	20 ^{b,c}	90.0%	6	15	21 ^c	80.9%	14	8	22	59.1%
	Included variables PC1, PC3, PC4												
	PC2, PC3, PC5, PC6, PC7												
1	Chi2	13.24; df = 1, p < 0.001			7.28; df = 1, p < 0.01				1.31; df = 1, p < 1				
C	Cpro	52.0% min. hit-rate = 65.0%			52.8% min. hit-rate = 66.0%				51.6% min. hit-rate = 64.5%				
	Cmax	60.0% min. hit-rate = 75.0%			61.9% min. hit-rate = 77.3%				59.0% min. hit-rate = 73.7%				

^a Cross validation employed: leave out 1 row in turn. ^b Entry excluded: 3. ^c Entry excluded: 11

Table 6. NNRTIs classification into classes **R** and **S**

Entry	Descriptors				SHAPE			
	L100I	K103N	V106A	Y181C	L100I	K103N	V106A	Y181C
1	R	R	s	S	R	R		S
2	r	r	r	S	s		r	S
3	EXCLUDED					R		R
4		s	s			r		
5			r	r				
6			r	r				
7		s	s			r		r
8	R	R		S	R	R		S
9	s		r	r			r	
10		s	r					s
11	EXCLUDED				EXCLUDED			
12	R	R		S	R	R		S
13		s					r	
14	s	r	s	s	EXCLUDED			
15	S	R	R	R	S	R	R	R
16	R	R	s	S	R	R		S
17		r		r				
18		S		r		s		
19		s	r	r		r	R	r
20	R	R	s	S	r	r	s	s
21		R		R		R	r	R
22	s	R				R	r	r
23	r	r		s	r	r		s
24	s		s	s			r	
25	S	s	r	R	s	s		R
26	s		r	r				
27		S		R		S		R
28		S		R		S		R
29		R		R		r	r	R
30	s		r	r				
31	S	r	R	r	s	r	R	r
32	s	r		r				
33		s	s					
34		s	s					
35		s					r	s
36	s	S	S		S	S	S	
37	s	S	S		S	S	S	
38		s	r					
39	r	s	s	s	EXCLUDED			
40	s			r		s		R
41	S	s		s	S	s		S
42		s	s			r		
43	S	s	S		S	S	S	
44	S	s	S		S	S	S	
45	S	R		r	S	r		r
46	s		r				r	
47	r	r	s	s				
48	s							
49	s							
50		R	s			R	r	
51	r	R		r		R	r	R
52	S	R		R	S	R	s	r
53	R	r	s	s	r	R		S
54	r	r	r	R	r			R
55	s		s					

R, S: correct; r, s: predicted; **R**, **S**: misclassified opposite class; **r**, **s**: misclassified no-data class

On the basis of the above results new calculations were carried out also including this time the derivatives for which data against mutant are not available and excluding entries exhibiting missing values. The classification procedure, in the present case, implies the fitting of compounds unclassified against each RT mutant into the classes **S** (susceptible) or **R** (resistant). Data reported in Table 6 show that it was possible to provide predictions on the activity of several derivatives assigning up to 10 compounds to the class **R**, whereas it has to be expected that an even higher number of derivatives, classified as **S**, can maintain their activity against RT mutated strains. In general, the best results were obtained in the case of L100I and K103N mutants for which the higher number of assignments was found when the principal components derived from the molecular descriptors was used. As far as V106A mutation is concerned the predictions are less significant because of the small number of input data available.

As far as the compounds classification is concerned, our results can be summarized as follow:

1. Comparison of the results for all the RT mutants shows that an unique consistent classification as **S** was achieved in the case of derivatives **33**, **34**, **40**, **41**, **48**, **49**, **55** and as **R** in the case of compounds **5**, **6**, **17**, **51**, **54**. The following derivatives were generally classified as susceptible but assignment was controversial against K103N mutant, entries **4** and **42**, or against V106A, entry **24**.
2. Derivatives **23**, and **41** were always predicted to the same class.
3. With regard to the 3 approved drugs, Delavirdine (**15**) was generally correctly identified with the sole exception of shape similarity index and Y181C mutant; Efavirenz (**20**) was generally classified correctly by the PC derived from descriptors, whereas Nevirapine (**31**) was well classified in the case of V106A mutant.
4. In the case of PNU-142721, the only compound under clinical trial for which so far different inhibitory activity and resistance/susceptibility data are available for the two stereoisomers [entries **40** (**R** configuration) and **41** (**S** configuration)], the analyses involving the descriptors and the shape similarity index in most cases were able to discriminate between the two isomers. For the other chiral compounds included in this study no relevant differences were evidenced, confirming thus the literature results of biological tests carried out on the mixture of isomers.

4 Conclusions

On this basis the multivariate statistical procedure PCA can be proposed as a reliable method for the prediction of the activity of NNRTIs, for which the data against mutant strains has not been reported.

Although this method implies some simplification of a more complex problem (a wider or diverse choice of descriptors could be considered) the approach proposed by us can be used as a sufficiently good and fast discriminator to preliminarily evaluate the probable emergence of resistance to newer synthesized compounds before it is actually verified in biological tests. This *in silico* screening results not expensive and easily accessible to most of researchers active in the field.

References

- [1] E. De Clercq, Novel compounds in preclinical/early clinical development for the treatment of HIV infections, *Rev. Med. Virol.* **2000**, *10*, 255–277.
- [2] G. Hajos, S. Riedi, J. Molnar, D. Szabo, Nonnucleoside reverse transcriptase inhibitors, *Drugs Fut.* **2000**, *25*, 47–62.
- [3] web address: <http://www.niaid.nih.gov/daids/dtpdb>.
- [4] Data taken from the following papers and reference cited therein: a) E. De Clercq, Non-nucleoside reverse transcriptase inhibitors (NNRTIs) for the treatment of human immunodeficiency virus (HIV-1) infections: strategies to overcome drug resistance development, *Med. Res. Rev.* **1996**, *16*, 125–157; b) G. Campiani, A. Ramunno, G. Maga, V. Nacci, C. Fattorusso, B. Catalanotti, E. Morelli, E. Novellino, Non-nucleoside HIV-1 reverse transcriptase (RT) inhibitors: past, present, and future perspective, *Curr. Pharm. Design* **2002**, *8*, 615–657; c) J. Corbett, L. Gearhart, S. Ko, J. Rodgers, B. Cordova, R. Klabe, S. Erickson-Viitanen, Novel 2,2-dioxide-4,4-disubstituted-1,3-*H*-2,1,3-benzothiadiazines as non-nucleoside reverse transcriptase inhibitors, *Bioorg. Med. Chem. Lett.* **2000**, *10*, 193–195; d) E. Tramontano, G. Piras, J. Mellors, M. Putzolu, H. Bazmi, P. La Colla, Biochemical characterization of HIV-1 reverse transcriptases encoding mutations at amino acid residues 161 and 208 involved in resistance to phosphonofomate, *Biochem. Pharm.* **1998**, *56*, 1583–1589; e) Z. Xu, R. Buckheit Jr., T. Stup, M. Flavin, A. Khilevich, J. Rizzo, L. Lin, D. Zembower, *In vitro* anti-human immunodeficiency virus (HIV) activity of the chromanone derivative, 12-oxocalanolide A, a novel NNRTI, *Bioorg. Med. Chem. Lett.* **1998**, *8*, 2179–2184; f) J. Balzarini, Suppression of resistance to drugs targeted to human immunodeficiency virus reverse transcriptase by combination therapy, *Biochem. Pharm.* **1999**, *58*, 1–27; g) R. Silvestri, M. Artico, S. Massa, T. Marceddu, F. De Montis, P. La Colla, 1-[2-(Diphenylmethoxy)ethyl]-2-methyl-5-nitroimidazole: a potent lead for the design of the novel NNRTIs, *Bioorg. Med. Chem.* **2000**, *10*, 253–256; h) M. Arranz, Synthesis and anti-HIV activity of 1,1,3-trioxo-2*H*,4*H*-thieno[3,4-*e*][1,2,4]thiadiazines (TTDs): a new family of HIV-1 specific non-nucleoside reverse transcriptase inhibitors, *Bioorg. Med. Chem.* **1999**, *7*, 2811–2822; i) R. Buckheit, E. White, V. Fliakas-Boltz, J. Russell, T. Stup, T. Kinjerski, M. Osterling, A. Weigand, J. Bader, Unique anti-human immunodeficiency virus activities of the nonnucleoside reverse transcriptase inhibitors calanolide A, costatolide, and dihydrocostatolide, *Antimicrob. Agents Chemother.* **1999**, *43*, 1827–1834; l) D. Zembower, S. Liao, M. Flavin, Z. Xu, T. Stup, R. Buckheit, A. Khilevich, A. Mar, A. Sheinkman, Structural analogues of the calanolide anti-HIV agents modification of the trans-10,11-dimethyldihydropyran-12-ol ring (ring c), *J. Med. Chem.* **1997**, *40*, 1005–1017.

- [5] Data taken from the databases available at the following web addresses: (a) <http://www.mediscover.net/antiviralintro.cfm>; (b) <http://resdb.lanl.gov/ResistDB/default.htm>; (c) <http://hivdb.stanford.edu/>; (d) <http://www.hivresistanceweb.com/>.
- [6] All the calculations were run on a Silicon Graphics Indigo II workstation by using the softwares TSAR 3.2 (Tools for Structure Activity Relationships), VAMP 6.0, ASP 3.2 (Automated Similarity Packages) (Oxford Molecular-Accelrys). Molecular descriptors were derived according to the method and assumptions reported in TSAR 3.2 Reference Guide, Oxford Molecular Limited, 1998.
- [7] R. Carbo, L. Leyda, M. Arnau, How similar is a molecule to another? An electron density measure of similarity between two molecular structures. *Int. J. Quantum Chem.* **1980**, 17, 1185–1189.
- [8] Tables with full matrices of similarity indexes are available as supplementary material.
- [9] B. F. J. Manly, Multivariate statistical methods. A primer, Chapman and Hall, 1986, Chapter 7.
- [10] C. Huberty, Issues in the use and interpretation of discriminant analysis, *Psychol. Bull.* **1984**, 95, 156–171.
- [11] http://www.georgetown.edu/faculty/ballc/webtools/web_chi.html.
- [12] L. Bachelier, S. Jeffrey, G. Hanna, R. D'Aquila, L. Wallace, K. Logue, B. Cordova, K. Hertogs, B. Larder, R. Buckery, D. Baker, K. Gallagher, H. Scarnati, R. Tritch, C. Rizzo, Genotypic correlates of phenotypic resistance to efavirenz in virus isolates from patients failing nonnucleoside reverse transcriptase inhibitor therapy. *J. Virol.* **2001**, 75, 4999–5008.
- [13] C. J. Petropoulos, N. T. Parkin, K. L. Limoli, Y. S. Lie, T. Wrin, W. Huang, H. Tian, D. Smith, G. A. Winslow, D. J. Capon, J. M. Whitcomb, A novel phenotypic drug susceptibility assay for human immunodeficiency virus type 1. *Antimicrob. Agents Chemother.* **2000**, 44, 920–928.

Received on May 22, 2003; Accepted on October 27, 2003

Why Wait to Make Great Discoveries

When you can make them in an instant with
Wiley InterScience® Pay-Per-View and ArticleSelect™

Now you can have instant, full-text access to an extensive collection of journal articles or book chapters available on Wiley InterScience. With Pay-Per-View and ArticleSelect™, there's no limit to what you can discover...

ArticleSelect™ is a token-based service, providing access to full-text content from non-subscribed journals to existing institutional customers (EAL and BAL)

Pay-Per-View is available to any user, regardless of whether they hold a subscription with Wiley InterScience.

Benefits:

- Access online full-text content from journals and books that are outside your current library holdings
- Use it at home, on the road, from anywhere at any time
- Build an archive of articles and chapters targeted for your unique research needs
- Take advantage of our free profiled alerting service, the perfect companion to help you find specific articles in your field as soon as they're published
- Get what you need instantly, no waiting for document delivery
- Fast, easy, and secure online credit-card processing for Pay-Per-View downloads
- Special, cost-savings for EAL customers: whenever a customer spends tokens on a title equaling 115% of its subscription price, the customer is auto-subscribed for the year
- Access is instant and available for 24 hours

www.interscience.wiley.com

